POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Unsupervised Illegal Landfills Detection using Land Cover specific Autoencoders

Laurea Magistrale in Mathematical Engineering - Ingegneria Matematica

Author: Elena Musiari
Advisor: Prof. Giacomo Boracchi
Co-advisor: Luca Frittoli, Ph.D.
Academic year: 2022-2023

1. Introduction

Nowadays, detecting illegal landfills, often hidden and causing severe environmental hazards, has become a critical challenge. In this work, we address the problem of illegal landfill detection through satellite images as an anomaly detection and segmentation problem, where landfills represent anomalies in the image and images without waste constitute the normality. This work is part of the European PERIVALLON project [1], which aims to fight organised environmental crime by developing solutions based on artificial intelligence to detect and contrast such illegal activities.
Detecting illegal landfills in satellite images has been carried out mostly by human experts, hence it was inefficient and limited compared to a possible automated version. Deep neural networks can offer a more efficient and effective solution, building an automated landfill discovery tool capable of detecting anomalies that human observation might miss, by continuously scanning the ground with satellite images.
Traditional monitoring methods for such tasks are supervised classification and segmentation. Instead, we use an unsupervised method to find illegal waste, without being conditioned by the types of anomalies during training.
The main challenge tackled by this thesis is the great heterogeneity of land covers without illegal waste, implying an extensive variety in normal images. Instead of creating a general model for all the satellite images, our solution suggests developing a specific model for each group of land covers and using it to detect illegal landfills only on images including those land cover types.
Combining methodologies from the fields of anomaly detection and graph theory, we use spectral clustering on the land covers to define strongly related groups and we split images into subsets. On these sets, we apply Fully Convolutional Autoencoders, followed by morphological image operations, to produce more efficient and accurate identification of illegal landfills.

2. Problem formulation

In this section, we state the anomaly detection problem that we address in this thesis. We refer to images containing an illegal landfill as *anomalous*, while *normal* images do not contain them. We can consider an RGB image in input as a three dimensional matrix $X \in \mathbb{R}^{w \times h \times 3}$, where values are normalized between $[0, 1]$. Here w is the width and h is the height of the image. Our goal is to locate anomalous regions in the image X defining an *anomaly mask*:

$$\Omega_X(i,j) = \begin{cases} 0 & \text{if } X(i,j) \text{ is normal} \\ 1 & \text{if } X(i,j) \text{ is anomalous} \end{cases}, \quad (1)$$

with $X(i,j)$ the pixel at row i and column j in X. We face this anomaly detection problem using an *unsupervised* approach, i.e. using unlabeled data. During the training phase, we take into account only *normal* images $X_1, ..., X_n \in \mathcal{X}$, namely images not containing any illegal waste. Unsupervised training considering only a type of data is also called *semisupervised* approach.

3. Proposed solution

3.1. Autoencoders for Anomaly Detection

We address the anomaly detection problem using an *Autoencoder*, namely a Convolutional Neural Network that is trained with the objective of reconstructing the input as desired output [5]. Autoencoders consist of two main components: the encoder $\mathcal{E}$ and the decoder $\mathcal{D}$. $\mathcal{E}$ maps the input into a low-dimensional latent space, from which $\mathcal{D}$ reconstructs the input image $\bar{X} = \mathcal{D}(\mathcal{E}(X))$. Our idea is to train the encoder and the decoder using only normal images such that the difference between the original image and the reconstructed image is minimized. When the difference is below a certain threshold τ, the image pixel $X(i,j)$ is considered normal, above the threshold the pixel is considered anomalous, namely

$$\begin{cases} \ell(X(i,j), \mathcal{D}(\mathcal{E}(X(i,j)))) \leq \tau & \text{if normal} \\ \ell(X(i,j), \mathcal{D}(\mathcal{E}(X(i,j)))) > \tau & \text{if anomalous} \end{cases},$$

with $\ell(\cdot,\cdot)$ the Autoencoder reconstruction loss. However, since in this dataset normality is too varied, training the model on the entire set of normal images results in poor segmentation performance. Therefore, we split the images into smaller groups according to their land cover. Formally, we create N groups of images $\mathcal{X}_1, \dots \mathcal{X}_N \subseteq \mathcal{X}$ such that $\bigcup_{i=1}^{N} \mathcal{X}_i \approx \mathcal{X}$, with $\mathcal{X}_i$ set of images consistent from the land cover perspective.

3.2. Dataset

In order to develop this thesis, we worked with the AerialWaste dataset [4] that collects satellite images of Lombardia region under the control of ARPA agency. The latest version of this dataset is composed of 10977 satellite images taken by three different sources: AGEA Ortophotos, WorldView-3 and GoogleEarth. According to the type of source, the images have different GSD (Ground Sampling Distance), namely a different pixel resolution of 20 cm, 30 cm or 50 cm respectively. Such images are already split into training set (75% of the total number of images) and test set (25% of the total). Among the test set, a subset of 169 anomalous images is provided with segmentation masks in the standard COCO format. We decided to follow the division provided by the dataset authors, even though in the training we use only normal images, namely 5579 images of the training set.
Thanks to the dataset DUSAF 7.0 published in the Geoportal of Lombardia region [2], we were able to associate each image to one or (usually) more types of land covers in the form of numerical codes. Indeed, AerialWaste dataset has no annotations about the land cover of each image, forcing one to consider as *normal* a huge variety of images. Figure 1 illustrates images from different land cover types, (a), (b) and (c) are normal images, while (c) and (d) are anomalous images with their ground truth masks.

3.3. Methods

The idea of our solution is to create small groups of strongly connected land cover types, in which the concept of *normality* has limited variability. On these groups, we train an Autoencoder that will be specific for the land covers within the group. Then, at test time, we evaluate the performance of the specific Autoencoders on normal and anomalous test images belonging to the same group.
In order to select only a subset of normal images to focus the work on, we group the different types of land cover using the *spectral clustering* algorithm. To achieve this, we create an undirected weighted graph $G = (V, E)$ where each node $i \in V$ is a different type of land cover, and each edge $e_{ij} \in E$ connecting i and j has weight w_{ij} equal to the number of images containing simultaneously i and j as land covers. Thus, the weight on each edge is a function of *similarity* between the two vertices that the edge connects, namely it represents the similarity between two land cover types. Our aim is to highlight the connections between types of land cover that are strongly related, since a high edge weight corresponds to a high number of images where those types are co-present.
We are able to apply to this graph the Normalized minimal cut algorithm provided by Shi and Malik [3]: partition the graph G into disjoint sets of vertices $A_1, \dots, A_K$, i.e. such that

(a) 3121, 31311 (b) 1121, 1122, 1221, 12111, 1112 (c) 3113, 5121 (d) 12112, 2111 (e) 12112, 2111, 11231

Figure 1: Images (a), (b), (c) are *normal*: (a) has categories "medium and high density coniferous forests (3121)", "medium and high density mixed forests governed by coppice (31311)"; (b) has "discontinuous residential fabric (1121)", "sparse and nucleiform residential fabric (1122)", "road networks and ancillary spaces (1221)", "industrial, craft, commercial settlements (12111)", "medium dense continuous residential fabric (1112)"; (c) has "riparian formations (3113)", "natural water basins (5121)". Instead (d), (e) are *anomalous* images with their segmentation masks, the first given along with the dataset, the latter manually segmented by us: (d) and (e) have land covers "agricultural production facilities (12112)", "simple arable land (2111)"; (e) has also "farmhouses (11231)".

$\bigcup_{i=1}^{K} A_i = V$ and $\forall i, j \; i \neq j, \; A_i \cap A_j = \emptyset$, by removing edges connecting the two sets. The objective is to eliminate edges with a certain criterion minimizing the total weight of the edges deleted. The sum of the weights of the removed edges represents the degree of dissimilarity, and it is called *cut*: $cut(A_i, A_j) = \sum_{i \in A_i, \, j \in A_j} w_{ij}$. The *normalized cut (Ncut)* avoids cutting small sets of isolated nodes in the graph, and the cut cost is computed as a fraction of the total edge connections to all the nodes in the graph:

$$Ncut(A_i, A_j) = \frac{cut(A_i, A_j)}{tot(A_i, V)} + \frac{cut(A_i, A_j)}{tot(A_j, V)}$$

where $tot(A_i, V) = \sum_{l \in A_i, \, k \in V} w_{lk}$ is the total connection, i.e. the sum of the weights of edges from the vertices in A_i to all the other vertices. Since minimizing the normalized cut is an NP-complete problem, an approximate discrete solution can be efficiently found: relaxing the hypothesis to real values only, the problem can be rewritten as the solution of the generalized eigenvalue system

$$(\mathbf{D} - \mathbf{W})y = \lambda \mathbf{D} y \qquad (2)$$

where $\mathbf{D}$ is the $N \times N$ diagonal matrix with the total connection $d(i) = \sum_j w_{ij}$ from node i to all the other vertices on its diagonal, $\mathbf{W}$ is the $N \times N$ symmetrical matrix with the graph weights as elements $\mathbf{W}(i, j) = w_{ij}$. It is found in (2) that the second smallest eigenvector is the relaxed solution of the normalized cut problem, thus we use spectral clustering to partition the graph.
We perform this analysis on the training set, where there are only *normal* images, namely without illegal landfills, in order to catch possible relations between land cover types in a context of normality.

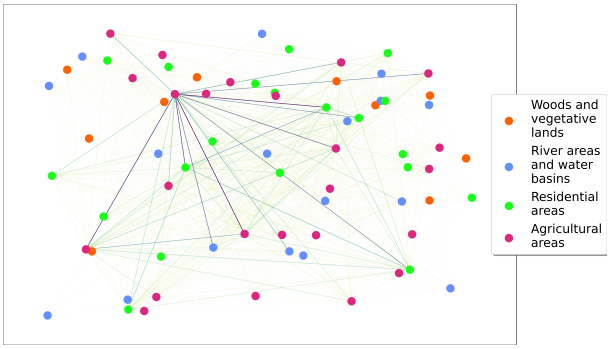


Figure 2: The clustered graph found with spectral clustering: each node i is a different land cover, each edge (i, j) has weight w_{ij} equal to the number of images containing simultaneously land covers i and j. The colour of each edge stresses the value of its weight (lighter colour, smaller weight). The clusterization represents the main groups of land covers we can find in Lombardia territory.

In order to choose the optimal number of clusters k, we look for the value k maximizing the eigengap, namely the difference between consecutive eigenvalues, which turns out to be $k = 4$. Performing the spectral clustering based on land covers, we discover 4 different groups, as shown in Figure 2, that can be summed up as:

1. Woods and vegetative lands
2. River areas and water basins
3. Residential areas
4. Agricultural areas.

Our analysis is developed on the Agricultural areas cluster, which contains the land cover types listed in Table 1.

Code	Description
223	Olive groves
31111	Medium and high density deciduous forests governed by coppice
2111	Simple arable crops
2311	Permanent meadows in the absence of tree and shrub species
2112	Arable land trees
12112	Agricultural production sites
3241	Bushes with significant presence of tall shrub and tree species
3242	Bushes in abandoned agricultural areas
1123	Sparse residential fabric
11231	Farmhouses
221	Vineyards
2312	Permanent meadows with scattered tree and shrub species
31122	Low density broad-leaved forests governed by high trunk
222	Orchards and minor fruit
21141	Open field floro-nursery crops
21142	Protected floro-nursery crops
21131	Vegetable crops in open field
21132	Protected horticultural crops
31121	Low density deciduous forests governed by coppice
31112	Medium and high density broad-leaved forests governed by high stems
12126	Photovoltaic systems on the ground
3114	Chestnut groves
3111	Medium and high density hardwood forests

Table 1: Codes and descriptions of the land covers in *Agricultural areas* cluster.

4. Implementation details

4.1. Preprocessing

Since the images have different pixel-resolution, we scale them according to their source choosing to set 30 cm/pixel for all of them. Even if the network is able to receive in input images of all sizes, in order to help the learning using batches we randomly crop patches from the training images and give them in input to the encoder. Keeping in mind the size of the receptive field of our network, we decide to crop the patches of size 128×128 pixels, namely 3840×3840 cm, and we create batches of 128 patches each.

4.2. Postprocessing

During the testing, error maps are created as reconstruction error of the images, using the training losses. These maps can be described as matrices $Z = \ell(X, \bar{X}) \in \mathbb{R}^{w \times h}$, with the values of Z high where the reconstruction fails.
In order to create a *score map*, we choose a threshold τ from the empirical quantile (at 98% and 99%) of the distribution of the error. Thereby, following Equation (1), the score map can be created in each pixel $X(i, j)$ as:

$$\Omega_X(i,j) = \begin{cases} 0 & \text{if } \ell(X(i,j), \bar{X}(i,j)) < \tau \\ 1 & \text{if } \ell(X(i,j), \bar{X}(i,j)) \geq \tau \end{cases}$$

with i, j indicating the row and column locating the pixel in the image, ℓ the reconstruction loss. The score map created is then postprocessed alternating the morphological methods of *erosion* and *dilation*. Erosion shrinks the shape of an object in the image by removing pixels from its edge, while dilation increases the shape by adding pixels. The amount of pixels changed in the image depends on the size and shape of the structuring element used to process the image. In such a way, the noise in the prediction map of the anomaly masks is removed, and the holes in the objects are filled.

4.3. Neural network training

In this thesis we consider three different losses to train the networks. The first is *MSE* that computes the *mean squared error* between each pixel of the two images taken in comparison, namely the original and the reconstructed image, and it is defined as

$$MSE(X, \bar{X}) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} (X(i,j) - \bar{X}(i,j))^2$$

where the difference between images is considered pixel-wise, $X(i, j)$ the pixel ranging in n rows and m columns.
Besides, the *structural similarity index* (*SSIM*) looks for the similarity between pixels of the two images, producing a score $\in \{-1, 1\}$, with 1 indicating the maximal similarity. It is defined as

$$SSIM(X, \bar{X}) = \frac{(2\mu_X \mu_{\bar{X}} + C_1)(2\sigma_{X\bar{X}} + C_2)}{(\mu_X^2 + \mu_{\bar{X}}^2 + C_1)(\sigma_X^2 + \sigma_{\bar{X}}^2 + C_2)}$$

where μ_X, $\mu_{\bar{X}}$ are the mean of X and $\bar{X}$, σ_X, $\sigma_{\bar{X}}$ the variance of X and $\bar{X}$, $\sigma_{X\bar{X}}$ their covariance, C_1, C_2 constants to avoid instability when the sum of squared means is close to 0. In order to use this metric during the training as a loss, we consider $1 - SSIM(X, \bar{X})$.
Finally, we define a mixed loss as

$$\ell_{mixed} = w_{MSE} \cdot MSE(X, \bar{X}) + w_{SSIM} \cdot (1 - SSIM(X, \bar{X})) \quad (3)$$

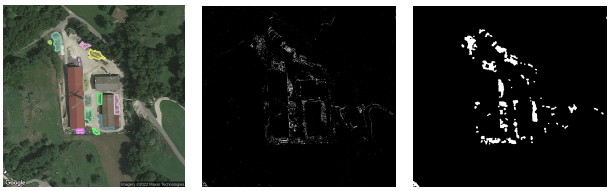
with w_{MSE}, w_{SSIM} weights arbitrarily chosen as hyperparameters. In this way we try to combine the MSE reconstruction precision with the SSIM ability to grasp the image structure.
Before performing the training, we further split the cluster of Agricultural areas into *frequent* and *rare*, where the first contains only the most frequent land cover types, namely those with at least 50 samples, and the latter contains the other more rare types. We obtain 1560 normal images in the frequent training set, 307 images (54 anomalous) in the rare test set, 787 images (285 anomalous) in the frequent test set. The aim is to use only the frequent subset in the training phase, and use the rare land cover types exclusively during the testing phase in order to check the generalization ability of the network.

4.4. Autoencoder Architecture

Since the images have different dimensions, we design a Fully Convolutional Autoencoder in order to take in input images of any size. Two different Fully Convolutional architectures are used: a network with 3 convolutional layers in the encoder, with ReLU as activation function, and in the decoder 3 convolutional layers alternated with upsampling layers, with ReLU as activation function; then a similar network with 4 layers instead of 3. The first network is trained with MSE loss, instead the second with the mixed loss between MSE and SSIM defined in (3) of weights $w_{MSE} = 0.7$, $w_{SSIM} = 0.3$.

5. Experiments

In this thesis we perform three different type of model evaluation: performance of the model using AUROC, ability to segment in binary masks, and ability to find a sufficient amount of waste in images. In this section we list all the results obtained during our evaluation experiments on the test set, that contains both normal and anomalous images. Since the number of segmented images in the dataset is scarce, especially if we consider only the images in our cluster, we manually segment approximately 100 other images using Odin annotator provided by the authors of AerialWaste dataset[4].
We first estimate the AUROC score on the reconstruction error compared to the ground truth masks. Then, as explained in Section 4.2, we determine the score maps for the anomalous test images of the Agricultural cluster and we compute the *IoU* and *Dice* scores of those predicted maps with respect to the ground truth masks. Initially we do not apply the morphological postprocessing, but we notice that applying it on the score maps the results improve significantly.



(a) (b) (c)

Figure 3: Image (a) is the GT mask, (b) is the mask computed through the model trained with MSE before postprocessing and has scores $IoU = 0.1509$, $Dice = 0.3017$, (c) after postprocessing has scores $IoU^{postpr} = 0.4704$, $Dice^{postpr} = 0.9408$.

Figure 3 shows an example of the difference between the raw predicted masks (b) and the mask after postprocessing (c).

	IoU	Dice	AUROC
Baseline	0.0628	0.1257	69.91 %
Baseline^{postpr}	0.1446	0.2893	
$\ell_{\mathbf{MSE}}$	0.0629	0.1257	70.12 %
$\ell_{\mathbf{MSE}}^{\mathbf{postpr}}$	0.1482	0.2964	
$\ell_{\mathbf{mixed}}$	0.0452	0.0903	70.99 %
$\ell_{\mathbf{mixed}}^{\mathbf{postpr}}$	0.0881	0.1762	

Table 2: Evaluation of the models trained with different losses, before and after the application of morphological postprocessing; as baseline we consider a model trained on the complete dataset.

In Table 2 are listed the results considering at 99% the threshold cutoff defined in Section 4.2, with and without applying postprocessing. We observe that after the postprocessing the performance improve in all our models; the mixed-loss model has best AUROC score but it is the worst considering IoU and Dice scores, where excels the MSE-based model. The main issue with these results is the false detection of object edges as anomalies, for instance roof edges, roadsides, small items. This matter is directly related to the problem type that we are addressing, that is at pixel-level.
As previously stated, we also perform the evaluation on the rare subset of the cluster in order to know if the models were able to generalize what learnt on the frequent subset. Performance is shown in Table 3 and is consistent with the results of the frequent subset.

	IoU	Dice	AUROC
$\ell_{\mathbf{MSE}}{}^{\mathbf{postpr}}$	0.15806	0.3161	78.01 %
$\ell_{\mathbf{mixed}}{}^{\mathbf{postpr}}$	0.0676	0.1352	77.51 %

Table 3: Evaluation on the rare set of cluster Agricultural areas of the models trained on frequent set with MSE loss and mixed loss.

Instead, evaluating the models on the images contained in another different cluster, more specifically on the Residential areas cluster, we obtain the results shown in Table 4.

	IoU	Dice	AUROC
$\ell_{\mathbf{MSE}}{}^{\mathbf{postpr}}$	0.1273	0.2547	68.94 %
$\ell_{\mathbf{mixed}}{}^{\mathbf{postpr}}$	0.0757	0.1515	68.33 %

Table 4: Models trained on cluster of Agricultural areas evaluated on cluster of Residential areas.

These results denote that our models can broaden what learnt on frequent land covers to connected and affine land cover types, but also that the mixed loss model is not specific for Agricultural cluster. Therefore, we use only the model built with MSE loss to evaluate the prediction of a sufficient amount of waste in anomalous images of the frequent subset.
Computing the value minimizing the distance between precision and recall scores, we find the optimal threshold for the IoU score. We use this threshold to discriminate between images predicted as *anomalous* (namely with a sufficient amount of waste detected) or *normal*, finding: 59 images properly predicted as anomalous (true positive), 13 incorrectly predicted (simultaneously false negative and false positive, since the masks were misplaced). Performing the same evaluation considering only normal images we find 62 true negative and 440 false positive.
Hence, we can complete the testing computing $Recall = 0.952$ and $Precision = 0.134$. Since in the real-world scenario the most important thing is not to miss any illegal landfill, the major result is that almost all the anomalies are correctly detected, as enhanced by the high recall value.

6. Conclusions

In this work we approached the illegal landfills detection problem from a different perspective with respect to the previous works, namely as a unsupervised anomaly detection problem.
Addressing the problem in an unsupervised pixel-level manner is less effective with respect to supervised image-level approaches, but it gives us the benefit of using the same network with new different images, without retraining it. Indeed, in a context such as illegal landfills detection, it is important to be flexible since the satellite image dataset could be in continuous update.
Clustering the land cover types slightly improved the performance of our models, even if the edges of objects in the image are likely to raise false alarms. This issue is due to the complexity of the addressed task and the use of MSE loss to train our model. We have shown that this approach is feasible and viable, bringing reasonable results.
The mixed loss turned out to be underperforming with respect to the classical MSE loss. As a future approach, it could be developed a mixed loss using a different type of structural similarity or a different metric. Overall, we noticed that the results depend also on the quality of the ground truth masks, since model predictions are pixel sensitive and their assessment is affected by that.
We underline that in our context it is more important to have a high recall with respect to a high precision, meaning that false alarms of illegal landfills could be risen, but almost all of the illegal areas are detected: sending someone to investigate a suspicious area is better than not spotting it.

References

[1] European Commission CORDIS. PERIVALLON Protecting the EuRopean terrItory from organised enVironmentAl crime through inteLLigent threat detectiON tools. `https://cordis.europa.eu/project/id/101073952`, 2022-2025.

[2] Geoportale della regione Lombardia. Uso e copertura del suolo 2021 (dusaf 7.0). `https://www.geoportale.regione.lombardia.it`, updated 2023.

[3] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[4] R. N. Torres and P. Fraternali. AerialWaste dataset for landfill discovery in aerial and satellite images. *Scientific Data*, 10(1):63, 2023.

[5] C. Zhou and R. Paffenroth. Anomaly detection with robust deep autoencoders. pages 665–674, 08 2017.